An Integrated Data Analytics Platform

Edward M. Armstrong[1], Mark A. Bourassa[2], Tom Cram[3], Jocelyn Elya[2], Frank R. Greguska III[1], *Thomas Huang[1], Joseph C. Jacob[1], Zaihua Ji[3], Yongyao Jiang[4], Yun Li[4], Nga Quach[1], Lewis McGibbney[1], Shawn Smith[2], Vardis Tsontos[1], Brian Wilson[1], Steve J. Worley[3], and Chaowei Yang[4]

[1] NASA Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA
[2] Center for Ocean-Atmospheric Prediction Studies, Tallahassee, FL, USA
[3] National Center for Atmospheric Research, Boulder, CO, USA
[4] George Mason University, Fairfax, VA, USA

Correspondence:
Thomas Huang
thomas.huang@jpl.nasa.gov

1. Abstract

An Integrated Science Data Analytics Platform is an environment that enables the confluence of resources for scientific investigation. It harmonizes data, tools and computational resources which subsequently enable the research community to focus on the investigation rather than spending time on security, data preparation, management, etc. OceanWorks is a NASA technology integration project to establish a cloud-based Integrated Ocean Science Data Analytics Platform at NASA's Physical Oceanography Distributed Active Archive Center (PO.DAAC) for big ocean science. It focuses on advancement and maturity by bringing together several NASA open-source, big data projects for parallel analytics, anomaly detection, in-situ to satellite data matchup, quality-screened data subsetting, search relevancy, and data discovery.

Our communities are relying on data distributed through data centers such as the PO.DAAC, COAPS, NCAR, and many others to conduct their research. In typical investigations, scientists would engage in: search for data, evaluate the relevance of that data, download it, and then apply algorithms to identify trends. Such workflow cannot scale if the research involves a massive amount of data or multi-variate measurements. NASA's Surface Water and Ocean Topography (SWOT) mission is expected to produce massive amount of observational data during its 3-year nominal mission. Collections like SWOT challenges all existing Earth Science data archival, distribution and analysis paradigms. In this paper, we will discuss how OceanWorks enhances the analysis of physical ocean data where the computation is done on an elastic cloud platform next to the archive to deliver fast, web-accessible services for working with oceanographic measurements.

2. Keywords

Big Data, Cloud Computing, Ocean Science, Data Analysis, Matchup, Anomaly Detection

3. Introduction

With increasing global temperature, warming of the ocean, and melting ice sheets and glaciers, the impacts can be observed from changes in anomalous ocean temperature and circulation patterns, to increasing extreme weather events and super hurricanes, sea level rise and storm

surges affecting coastlines, and may involve drastic changes and shifts in marine ecosystems. To date, investigative science requires researchers to work with many disjoint tools such as search, reprojection, visualization, subsetting, and statistical analysis. Researchers are finding themselves having to convert nomenclature between these tools, including something as mundane as dataset name and representation of geospatial coordinates. Sometime our researchers also required to transform the data into some common representation in order to coordinate measurements collected from different instruments. The concept of an Integrated Data Analytics Platform (Figure 1) is to tackle these data wrangling, management, and analysis challenges, so our researchers can focus on their investigation.

In recent years, NASA's Advanced Information Systems Technology (AIST) and Advancing Collaborating Connections for Earth System Science (ACCESS) programs have invested in developing new technologies targeting big ocean data on the cloud computing platform. Their goal is to address some of the big ocean science challenges by leverage modern computing infrastructure and horizontal-scale software methodologies. Rather than looking into developing a single ocean data analysis application, we have developed a data service platform to enable many analytic applications and lay the foundation for community-driven big ocean science.

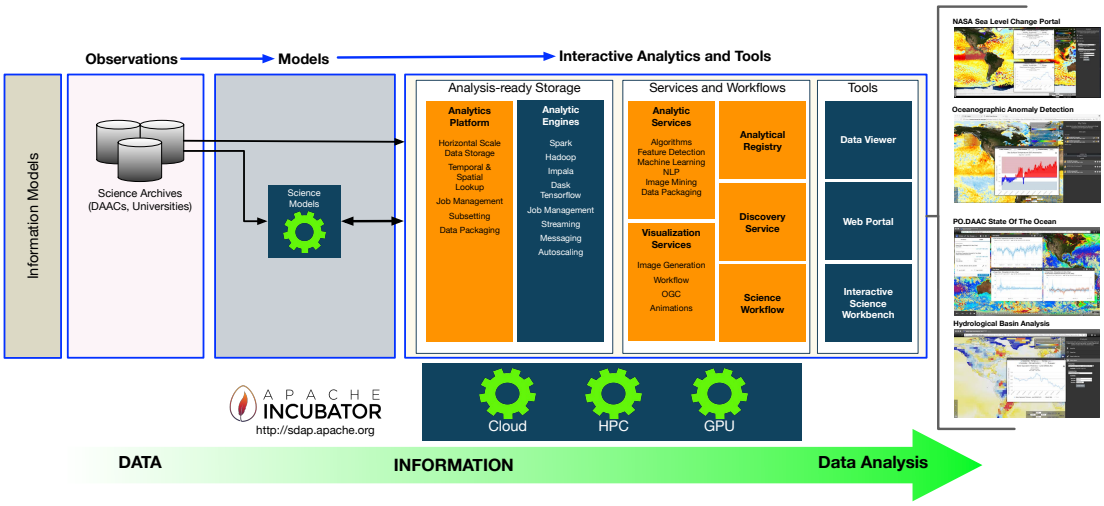


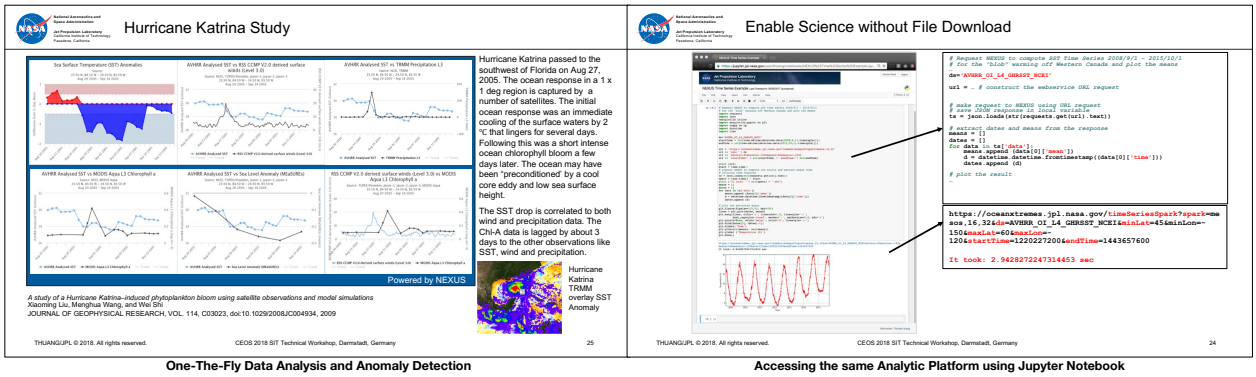
Figure 1: An Integrated Data Analytics Platform

OceanWorks (Huang, 2018) is a NASA AIST project to mature NASA's recent investments through integrated technologies and to provide the oceanographic community with a range of useful and advanced data manipulation and analytics capabilities. This 2-year development effort is to rollout an Integrated Data Analytic Platform for ocean science. This platform is designed to be extensible to promote community contribution with the following initial offerings:

- Data analysis
- Data-Intensive anomaly detection
- Distributed *in situ* to satellite matchup
- Search relevancy
- Quality-screened data subsetting

While the project is still in active development, in 2017 the OceanWorks project team has donated all of the project's source code to the Apache Software Foundation and established the official Science Data Analytics Platform (SDAP) project (http://sdap.apache.org) for community-driven and development of data access and analysis platform for the cloud environment. The OceanWorks project team is now develop in the open.

4. OceanWorks Components



One-The-Fly Data Analysis and Anomaly Detection **Accessing the same Analytic Platform using Jupyter Notebook**

Figure 2: Example OceanWorks Services

4.1 Data Analytics

We have been developing analytics solutions around common file packaging standards such as netCDF and HDF. We push for Climate Forecast (CF) recommendations and the Attribute Convention for Dataset Discovery (ACDD) to promote interoperability and improve our searches. Yet, there is very little effort in tackling our current big data analytic challenges, which includes how to work with petabyte-scale data and being able to quickly lookup the most relevant data for a given research. While the current method in subsetting analyzing one daily global observational file at time is the most straightforward, it is an unsustainable approach for analyzing petabyte of data. The common bottleneck is in working with large collection of files. Since these are global files, researchers are finding themselves having to move (or copy) more data than they need for their regional analysis. Webservice solutions such as OPeNDAP and THREDDS provide a webservice API to work with these data, but their implementation still involves iterating through large collection of files.

The OceanWorks' analytics engine is called NEXUS. It takes on a different approach for storing and analyze large collection of geospatial, array-based data by breaking the netCDF/HDF file data into data tiles and store them into a cloud-scale data management system. With each data tile has its own geospatial index, a regional subset operation only requires to retrieve the relevant tiles into the analytic engine. Our recent benchmark shows NEXUS can compute an area-averaged time series hundreds time faster than traditional file-based approach (Jacob, 2017).

OceanWorks enables advanced analytics that can easily scale to the available computation hardware along the full spectrum from an ordinary laptop or desktop computer, to a multi-node server class cluster computer, to a private or public cloud computer. The architectural drivers are:

- Both REST and Python API interfaces to the analytics
- In-memory map-reduce style of computation

- Horizontal scaling so computational resources can be added or removed on demand
- Rapid access to data tiles that form natural spatiotemporal partition boundaries for parallelization
- Computation performed close to the data store to minimize network traffic
- Container-based deployment

The REST and Python API enables OceanWorks to be easily plugged into a variety of web-based user interfaces, each tuned to particular domains. Calls to OceanWorks from a Jupyter notebook enables interactive cloud-scale, science-grade analytics.

Built-in analytics are provided for the following algorithms:

- Area-averaged time series to compute statistics (e.g., mean, minimum, maximum, standard deviation) of a single variable or two variables being compared. Optionally apply seasonal or low-pass filter to the result.
- Time-averaged map to produce a geospatial map that averages gridded measurements over time at each grid coordinate within a user-defined spatiotemporal bounding box.
- Correlation map to compute the correlation coefficient at each grid coordinate within a user-specified spatiotemporal bounding box for two identically gridded datasets.
- Climatological map to compute a monthly climatology for a user-specified month and year range.
- Daily difference average to subtract a dataset from its climatology, then, for each timestamp, average the pixel-by-pixel differences within a user-specified spatiotemporal bounding box.
- In-situ match to discover in situ measurements that correspond to a gridded satellite measurement.

In addition, authenticated or trusted users may inject their own custom algorithm code for execution within OceanWorks. An API is provided to pass the custom code as either a single or multi-line string or a python file or module

4.2 *In Situ* to Satellite Matchup

Comparison of measurements from different ocean observing systems is a frequently used method to assess the quality and accuracy of the measurements. The matching or collocating and evaluation of *in situ* and satellite measurements is a particularly value method because the physical characteristics of the observing systems are so different and therefore the errors related to instrumentation and sampling are not convoluted. The satellite community tends to use collocated *in situ* measurements to develop, improve, calibrate, and validate the integrity of the retrieval algorithms. The *in situ* observational community uses collocated satellite data to assess the quality of extreme/suspicious values and to add spatial context to the often spares point values. In both of these research realms there are many more detailed use cases not mentioned here, e.g. near real time decision support of field programs, planning exercises for future observing system deployments, and development of integrated, *in situ* plus satellite data, global gridded analyses products that are useful for stand-alone research and for model initialization and boundary conditions.

There are several major data challenges related to successful satellite and *in situ* data collation research. Disparate data volume and variety is primary challenge. Individual satellite

collections are typically large in volume, have relatively homogenous sampling, are derived from a single platform, are composed of a consistent set of parameters, and are represented as scan lines, swaths, or globally gridded fields. *In situ* observations typically bring the variety challenge into the problem. They are often replete with heterogeneous observing platforms (ships, buoys, Argo, glides, and etc.), instrumentation types and sampling methods, highly varying sampling rates, isolated spatio-temporal coverage over the global ocean. Another major challenge for collation-based research is logistical. The archives of satellite data and *in situ* data are often distributed at different centers, have a variety of access methods that need to be understood and applied, have data formatting and quality control information that are different, and over time the data can be dynamically extend (adding data to the time series) and completely new versions with critical data quality improvements can be made available. The OceanWorks matchup service (Smith et al. 2018) resolves these major challenges and many other secondary challenges.

4.3 Quality-Screened Subsetting

When working with earth science data and information, whether derived from an *in situ* platform, or airborne or satellite instruments, users often need to access, understand and apply data quality information, such as those contained in variables that specify quality information related to instrument and algorithm performance, and other synoptic environmental characteristics or conditions. The ability to screen the physical data records via services that apply standardized sets of quality flags, states or conditions is imperative to allow scientists to seamless use these data to meet their requirements for error and accuracy, and many other use cases.

In the oceanographic *in situ* realm there are a number of models and conventions in use by the community. The Oceanworks project has chosen the IODE (International Oceanographic Data and Information Exchange) convention, an internationally recognized and developed approach to screen *in situ* observations based on both a primary and secondary level of flagging schemes. We have chosen only the five-level primary scheme because of its simplicity and straightforwardness allows a direct mapping and transformation to the native quality flags embedded in our ICOADS and SAMOS *in situ* datasets.

In the oceanographic satellite realm, a similar need for standardization exists that is exasperated by the increasingly denser availability of quality information in the form of data accuracy, processing algorithms states and failures, environmental conditions, and auxiliary variables that are packed as conditions into quality variables represented as scalar flags or bit flags. This level of complexity makes it often difficult and confusing for a science user to understand and apply the proper flags to screen for meaningful physical data. The NASA software project, the Virtual Quality Screening Service (VQSS) (Armstrong et al, 2016), addressed these issues by implementing a service infrastructure to expose, apply, and extract quality screening information through implementations of strategic databases and web services, data discovery, and exposure of granule-based quality information interactive menus. Fundamentally it leveraged on the availability of Climate and Forecast (CF) metadata conventions applied to the satellite quality variables that strictly standardizes the structure and content of quality information through its attributes: *flag_values, flag_mask, and flag_meanings.* Web services employed were able to seamlessly extract physical information in the form of netCDF and JSON outputs based on screening conditions using these bit flag and scaler conditions, auxiliary variable for data threshold conditions, and many other use cases. Oceanworks will employ this architecture to allow users a similar capability to apply the quality information embedded in the gridded and

ungridded input satellite data sources for sea surface temperature, ocean color, sea level, wind and precipitation parameters.

4.4 Search Relevancy and Discovery

Retrieving appropriate datasets is the prerequisite for data analysis, however, as the geographic data increases faster than ever, it poses grant challenge for researchers and developers to efficiently access the desired data. The NASA Physical Oceanography Distributed Active Archive Center (PO.DAAC) supplies the Earth science community with massive Oceanography data observed by over 30 satellites and missions. Although the PO.DAAC portal provide valuable data service to facilitate searching process, it still has some limitations including 1) most geospatial the keyword-based search method is popular in Geospatial portals, which doesn't take semantic meaning of the query into account, for example, the search engine can't retrieve metadata only containing "SLP" for a query "sea level pressure"; 2) Only single attribute is used in the ranking algorithm in most geospatial portals, such as spatial resolution, processing level, monthly popularity. However, multidimensional preferences should be considered in the ranking process; 3) The data portals lack of data relevancy. Some quite useful data exist in the portal, but users don't know they are there. Data relevancy could help navigate users to the data they are searching for.

OceanWorks is equipped with a data discovery engine with a profile analyzer (Jiang et al, 2017), a knowledge base, a smart engine. Raw web usages logs are collected from multiple servers like HTTP server and FTP server and grouped into sessions through the profile analyzer. Reconstructed sessions make it possible to learn user history search behavior and clickstream data, which is a valuable source of learning vocabulary linkages in addition to metadata. A RankSVM model (Joachims 2002) is trained on a few predefined ranking features with optimal ranking list provided by domain experts, aiming to increase the rank of data more relevant to the query. A recommender calculates the relevancy between metadata using their attributes and logs. A knowledge base is populated to store information like domain terms linkages, metadata relevancy, as well as pretrained model for ranking and recommendation. When a user input a query in the search box, highly related terms will be extracted from the knowledge base to expand the original search query and the search engine will retrieve data using the rewritten query instead of the input query, resulting in a higher recall score. The retrieve datasets will not be displayed to the user directly but reranked by the pretrained model to achieve a better ranking list. If the user chooses to view a metadata, the recommender will retrieve a list of related datasets to date being viewed, helping the user efficiently find data he/she needs. In summary, the optimal workflow tries its best to help data consumer to acquire dataset efficiently and accurately with advanced data learning methods.

4.5 Supported Datasets

The list of datasets supported by OceanWorks has grown in the past two years. Below is the current list of supported datasets:

Atmosphere

- MODIS Aqua Daily L3 Atmospheres, Collection 6, variable Aerosol Optical Depth 550 nm (Dark Target) (MOD08_D3v6)
- MODIS Terra Daily L3 Atmospheres, Collection 6, variable Aerosol Optical Depth 550 nm (Dark Target) MOD08_D3v6)

- MODIS Aqua Monthly L3 Atmospheres, Collection 6, variable Aerosol Optical Depth 550 nm (Dark Target) (MOD08_D3v6)
- MODIS Terra Monthly L3 Atmospheres, Collection 6, variable Aerosol Optical Depth 550 nm (Dark Target) MOD08_D3v6)

Chlorophyll

- MODIS Aqua Level 3 Global Daily Mapped 4 km Chlorophyll a

Estimating the Circulation and Climate of the Ocean (ECCO)

- Monthly Mean Version 4 release 2 – Net Surface Fresh-Water Flux, Net Surface Heat Flux, Mixed-Layer Depth, Bottom Pressure, SEAICE Fractional Ice-Covered Area, Free Surface Height Anomaly, SEAICE Effective Snow Thickness, Total Heat Flux, Total Salt Flux
- Monthly Mean Version 4 release 1 – Net Surface Fresh-Water Flux, Net Surface Heat Flux, Mixed-Layer Depth, Ocean Bottom Pressure, SEAICE Fractional Ice-Covered Area, Free Surface Height Anomaly, SEAICE Effective Snow Thickness, Actual Sublimation Freshwater Flux, Total Heat Flux, Total Salt Flux

Gravity

- Center for Space Research (CSR) GRACE RL05 Mascon Solutions
- JPL GRACE Mascon Ocean, Ice, and Hydrology Equivalent Water Height RL05M.1 CRI filtered Version 2

Ocean Temperature

- GHRSST Level 4 MUR Global Foundation Sea Surface Temperature Analysis (v4.1)
- GHRSST Level 4 AVHRR_OI Global Blended Sea Surface Temperature Analysis (GDS version 2) from NCEI
- MODIS Aqua Level 3 SST Thermal IR Daily 4km Nighttime v2014.0
- MODIS Aqua Level 3 SST Thermal IR Daily 4km Daytime v2014.0

Salinity

- JPL SMAP Level 2B CAP Sea Surface Salinity V2.0 Validated Dataset
- JPL SMAP Level 3 CAP Sea Surface Salinity Standard Mapped Image Monthly V3.0 Validated Dataset

Sea Surface Height Anomalies (SSHA)

- JPL MEaSUREs Gridded Sea Surface Height Anomalies Version 1609

Wind

- Cross-Calibrated Multi-Platform Ocean Surface Wind Vector L3.0 First-Look Analyses

Precipitation

- TRMM (TMPA) Precipitation L3 1 day 0.25 degree x 0.25 degree V7 (TRMM_3B42_Daily) at GES DIS

- TRMM (TMPA-RT) Precipitation L3 1 day 0.25 degree x 0.25 degree V7 (TRMM_3B42_RT) at GES DISC

In Situ

- Shipboard Automated Meteorological and Oceanographic System (SAMOS)
- International Comprehensive Ocean-Atmosphere Data Set (ICOADS) Release 3, Individual Observations
- Salinity Process in the Upper Ocean Regional Study – 1 (SPURS1)
- Salinity Process in the Upper Ocean Regional Study – 2 (SPURS2)
- Global gridded NetCDF Argo only dataset produced by optimal interpolation (salinity variables)
- Global gridded NetCDF Argo only dataset produced by optimal interpolation (temperature variables)]

5. Applications and Infusion

The OceanWorks has been deployed for use by a number of NASA projects. Some of these include the NASA Sea Level Change Portal (SLCP), the GRACE Science Portal, and work is currently underway to integrate it with the State of the Ocean (SOTO) tool as part of the NASA PO.DAAC. Each project has slightly different needs, but all of them were able to utilize OceanWorks to fulfill their requirements.

The NASA SLCP contains a wealth of information about how the Earth's sea level is changing. It acts as a one stop shop for everything from news articles to data analysis. OceanWorks has been deployed as the engine behind the Data Analysis Tool that is part of the portal. The Data Analysis Tool focuses on providing fast and easy to use data analysis on a curated list of datasets that are important to the understanding of sea level change. Because OceanWorks is able to be deployed in many configurations depending on project requirements, it was a perfect fit for providing the data analysis capabilities required by SLCP. In this particular instance, only a single instance of OceanWorks was required to power the analysis because the datasets being analyzed are limited in resolution and frequency. This allows for real-time interactive analysis through the JavaScript front-end.

Similar to the NASA SLCP, the GRACE Science portal has limited requirements with respect to the amount of data that needs to be analyzed. However, this project needed to be able to run on public cloud infrastructure. So, while the user interface and data are similar in nature, the backend server is hosted using Amazon Web Services (AWS). This is possible because OceanWorks provides the flexibility to be deployed on a laptop, a single server, a bare metal cluster, or on a public cloud.

The NASA PO.DAAC deployment has different requirements from both SLCP and GRACE. The datasets hosted by PO.DAAC are very large and cover a wide time period. In order to provide analysis capabilities for these larger datasets, more than one server is needed for analysis. OceanWorks was built for this situation and can utilize Apache Spark to scale horizontally and spread the compute requirements across a cluster of machines. With this cluster setup, OceanWorks is able to handle the analysis of larger, more dense datasets.

The multiple deployments of OceanWorks have proven that it is capable of handling a wide range of requirements and deployment scenarios. From single node to multi node, on premise to on cloud, and small data to big data, OceanWorks can handle it all.

6. Challenges and Outlook

The Apache Science Data Analytics Platform (SDAP) is the open source implementation of OceanWorks. The project team recognizes it will take years of collaborative effort to create a big data solution that satisfies the needs from various science disciplines. OceanWorks demonstrated how to create a community-driven technology through well-managed open source development process. Unlike many emerging Earth Science big data solutions, SDAP is designed as a platform with simple RESTful API that supports clients developed in any programming language. This façade-based architectural approach enables SDAP to continue to evolve and leverage any new open source big data technology. OceanWorks only addressed some of the ocean science needs. It requires contributions from our community to help continue to evolve this open source technology. This project team would like this community to image having a common ocean analytic engine next to our distributed archives of ocean artifacts. Researchers or tools developers can interact with any of these analytics services, managed by the data centers, without having to move massive amount of data over the Internet.

7. Acknowledgement

The research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, in collaboration with Center for Ocean-Atmospheric Prediction Studies (COAPS) at the Florida State University, National Center for Atmospheric Research (NCAR), and George Mason University (GMU), under a contract with the National Aeronautics and Space Administration. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsement by the United States Government or the Jet Propulsion Laboratory, California Institute of Technology.

8. References

Armstrong, E.M, Z. Xing, C. Fry, S.J.S. Khalsa, T. Huang, G. Chen, T. Chin, and C. Alarcon. (2016): “A service for the application of data quality information to NASA earth science satellite records,” Proceeding of the 2016 American Geophysical Union Fall Meeting, San Francisco, CA, 2016.

Huang, T. (2018): “High Performance Open Source Platform for Ocean Sciences,” Proceeding of the 2018 Ocean Sciences Meeting, Portland, OR, 2018.

Jacob, J., F. Greguska, T. Huang, N. Quach, and B. Wilson. (2017): “Design Patterns to Achieve 300x Speedup for Oceanographic Analytics in the Cloud,” Proceeding of the 2017 American Geophysical Union Fall Meeting, New Orleans, LA, 2017.

Jiang, Y., Y. Li, C. Yang, and F. Hu. (2017): “Towards intelligent geospatial data discovery: a machine learning framework for search ranking,” Proceeding of the International Journal of Digital Earth, September 2017.

Joachims, T. (2002): “Optimizing Search Engines using Clickthrough Data,” Proceedings of the ACM Conference of Knowledge Discovery and Data Mining, 2002.

Smith, S.R., J.L. Elya, M.A. Bourassa, T. Huang, V.M. Tsontos, B. Holt, F.R. Greguska, N. Quach, S.J. Worley, and Z. Ji. (2018): "Integrating the Distributed Oceanographic Match-Up Service into OceanWorks," Proceedings of the 2018 Ocean Sciences Meeting, Portland, OR, 2018.